Genetic and selective constraints on the optimization of gene product diversity

Daohan Jiang[1], Nevraj Kejiou[2], Yi Qiu[2], Alexander F. Palazzo[2,*] & Matt Pennell[1,3,*]

[1]*Department of Quantitative and Computational Biology, University of Southern California, USA*
[2]*Department of Biochemistry, University of Toronto, Canada*
[3]*Department of Biological Sciences, University of Southern California, USA*
[*]Corresponding authors: alex.pallazo@utoronto.edu, mpennell@usc.edu

Abstract

RNA and protein expressed from the same gene can have diverse isoforms due to various post-transcriptional and post-translational modifications. For the vast majority of alternative isoforms, It is unknown whether they are adaptive or simply biological noise. As we cannot experimentally probe the function of each isoform, we can ask whether the distribution of isoforms across genes and across species is consistent with expectations from different evolutionary processes. However, there is currently no theoretical framework that can generate such predictions. To address this, we developed a mathematical model where isoform abundances are determined collectively by *cis*-acting loci, *trans*-acting factors, gene expression levels, and isoform decay rates to predict isoform abundance distributions across species and genes in the face of mutation, genetic drift, and selection. We found that factors beyond selection, such as effective population size and the number of *cis*-acting loci, significantly influence evolutionary outcomes. Notably, suboptimal phenotypes are more likely to evolve when the population is small and/or when the number of *cis*-loci is large. We also explored scenarios where modification processes have both beneficial and detrimental effects, revealing a non-monotonic relationship between effective population size and optimization, demonstrating how opposing selection pressures on *cis*- and *trans*-acting loci can constrain the optimization of gene product diversity. As a demonstration of the power of our theory, we compared the expected distribution of A-to-I RNA editing levels in coleoids and found this to be largely consistent with non-adaptive explanations.

Keywords: gene product diversity; post-transcriptional modification; evolutionary theory; optimization; constraint

Introduction

Different RNA and protein isoforms can be expressed from the same gene, resulting in a phenomenon known as gene product diversity [1]. A variety of processes can generate gene product diversity, such as alternative transcription initiation [2, 3], alternative splicing [4–8], alternative polyadenylation [9], post-transcriptional RNA modifications [10–13], alternative translation initiation [14], post-translational modifications [8, 15], and errors during RNA or protein synthesis [16–19]. The growing body of transcriptomic and proteomic data has unveiled substantial gene product diversity produced by different processes in diverse taxa, but functional significance of the alternative isoforms remains largely unknown [1, 7, 8, 12, 13].

One explanation for observed gene product diversity is the adaptive hypothesis that the alternative isoforms perform essential functions and are beneficial to the organism [19–21]. Cases of beneficial gene product modifications have been documented in various taxa. Notable examples of potentially adaptive modification events include a nonsynonymous A-to-I RNA editing event in a potassium channel protein that confers cold tolerance in polar octopuses [22], A-to-I editing events in filamentous fungi that fix premature stop codons in proteins involved in sexual reproduction [23, 24], alternative splicing of *Sxl* transcripts that regulate sex determination in dipteran insects [25], and some circular RNA isoforms that function as micro RNA in sponges [26, 27]. However, such cases collectively comprise only a small portion of known gene product diversity.

An alternative view suggests that gene product diversity is largely non-adaptive and reflect errors in biochemical processes. Gene product modification processes that result in gene product diversity, like all other biochemical reactions, are fundamentally stochastic and thus prone to errors. While natural selection can act to reduce the error rate, optimization will be limited by a drift barrier: at small effective population sizes, molecular errors with mild fitness effect cannot be purged efficiently by selection in the face of strong genetic drift and/or mutational pressure [28, 29]. This view has been supported by analyses of various types of gene product diversity, such as alternative splicing [30–33], alternative polyadenylation [34], A-to-I RNA editing [35–37], and C-to-U RNA editing [38]. It is also plausible that different isoforms of a gene's product are functionally equivalent, in which case the diversity *per se* is not adaptive even if the process that generates diversity is. That is, it is the amount of modification in a molecule rather than the precise location of any modification that matters. Processes that can potentially generate such neutral diversity include

N6-methyladenosine (m6A) modification of RNA [38–40] and protein phosphorylation [41, 42].

Furthermore, a machinery that generates gene product diversity can be maintained by making otherwise strongly deleterious mutations reasonably benign. In such a case, it could become indispensable over time as more such mutations are permitted and fixed, a process, known as entrenchment or "constructive neutral evolution" [43–46]. For example, A-to-I editing can "permit" G-to-A mutations as inosine (I) is recognized as guanine (G) during translation; this harm-permitting effect has likely contributed to maintenance of high A-to-I editing activity in coleoid cephalopods (subclass Coleoidea, including octopuses, squids, and cuttlefishes) [36]. Similarly, high C-to-U editing in plant organelles may have been entrenched after permitting T-to-C mutations [47–49].

One possible way to distinguish these alternative hypotheses in the absence of functional information for the vast majority of isoforms is to compare the observed gene product diversity within and between species to that expected under various evolutionary scenarios. However, such comparisons are not currently possible as we lack a theoretical basis for generating such expectations. While phylogenetic comparative methods have recently been applied to molecular phenotypes like gene expression levels (e.g., [50–54]), it is unclear whether conventional trait evolution models used in phylogenetic comparative analyses are suitable for modeling gene product diversity. To address these, we developed a mathematical model that connects patterns of variation in gene product diversity and the underlying evolutionary processes. In particular, we investigated two types of gene product modification processes that represent a broad range of processes that generate gene product diversity. The first type of modification simply converts an unmodified isoform to modified isoform(s) that can potentially be dysfunctional and/or toxic (Fig. 1A). Such modifications are not universally required for gene products to carry out their primary functions. Prime examples of such modifications include a variety of post-transcriptional RNA editing processes, where the RNA molecule is enzymatically modified into an alternative isoform [10–13]. Thus, we will refer to this type of processes as "editing-type". The second type of gene product modification process is required to produce the functional isoform, but can potentially produce mis-processed isoforms that could be dysfunctional and/or toxic (Fig. 1B). This class of modification is exemplified by RNA splicing in eukaryotes, which is generally required but can potentially produce toxic mis-spliced isoforms [4, 5]. Thus, this second type of gene product modification is referred to as "splicing-type". In both cases, each gene product modification event is regulated by a set of *cis*-loci and a *trans*-factor. Each *cis*-locus only affects a specific modification event and thus has a local

effect, whereas the *trans*-factor globally affects many modification events.

Under our model, we derived phylogenetic means of the modification level (i.e., relative abundance of modified isoforms) under different conditions, demonstrating how modification level is shaped by mutational pressure, genetic drift, and selection. We also investigated how opposing selection on the modification process shapes the coevolution of *cis*- and *trans*-acting loci underlying modification. At last, using computer simulations, we demonstrated that our model can recapitulate distribution of A-to-I RNA editing levels observed in empirical studies.

Results and Discussion

Modeling genetic architecture of isoform abundances

Under a simple model where an unmodified isoform, I_0, is converted to a modified isoform, I_1, rates at which their abundances in the cell changes over time can be written as

$$\begin{cases} \frac{dP_0}{dt} = \alpha - \beta P_0 - \gamma_0 P_0 \\ \frac{dP_1}{dt} = \beta P_0 - \gamma_1 P_1. \end{cases} \tag{1}$$

Here, P_0 and P_1 are abundances of I_0 and I_1, respectively, α is the rate at which I_0 is produced, β is the per-molecule net rate at which I_0 is converted to I_1, and γ_0 and γ_1 are I_0 and I_1's respective decay rates. An equilibrium is reached when both rates are equal to zero:

$$\begin{cases} \alpha - \beta P_0 - \gamma_0 P_0 = 0 \\ \beta P_0 - \gamma_1 P_1 = 0. \end{cases}$$

Solving the system of equations gives equilibrium isoform abundances:

$$\begin{cases} P_0 = \frac{\alpha}{\beta + \gamma_0} \\ P_1 = \frac{\alpha\beta}{\gamma_1(\beta + \gamma_0)}. \end{cases} \tag{2}$$

The same modeling approach can be generally applied to systems with more isoforms—e.g., I_0 is converted to more than one modified isoforms (see Methods).

In our model, the per-molecule conversion rate β is controlled by a *trans*-factor (i.e., enzyme that performs gene product modification) and a set of *cis*-loci (i.e., sequence motif modulating enzyme binding affinity). The *trans*-factor's effect on β is characterized by a *trans*-genotypic value, Q, which reflects the modification enzyme's expression level and/or catalysis efficiency. The *cis*-genotype's effect is summarized by a normalized *cis*-genotypic value $\hat{v}$. A high $\hat{v}$ indicates strong binding between the modification enzyme and the substrate, which results in high modification efficiency, whereas a low $\hat{v}$ means weak enzyme-substrate binding and low modification efficiency. Each *cis*-locus can have either an effector allele that facilitates enzyme binding, or a null allele that has no effect. In this study, we focused on a simple model where all loci's effector alleles have an equal, additive effect [29], so $\hat{v}$ is calculated as $\hat{v} = v/l$, where l is the number of *cis*-loci that affect the modification and v is the total number of effector alleles. While this study is focused on such a simple model, it can readily be extended to incorporate variation in different loci's contribution–for example, a skewed distribution where one locus has major effect while others' effects are much weaker.

Given values of Q and $\hat{v}$, β is calculated as

$$\beta = Q(C\hat{v} + \epsilon). \tag{3}$$

Here, $C > 0$ represents whole-molecule features that modulate the *cis*-loci's effect size (e.g., secondary structure of RNA or protein), and $\epsilon \geq 0$ is the rate of non-specific modification (i.e., promiscuous activity of the enzyme independent of the *cis*-genotype).

For editing-type modification, we focused on a simple scenario where two isoforms, the unmodified isoform I_0 and modified isoform I_1, are present (Fig. 1A); the generic, two-isoform model described above is thus readily applicable. We considered values of l that are relatively small (i.e., no more than 10), as empirical studies suggest that sequence motifs with major effects on RNA modifications usually consist of a small number of nucleotide sites [10, 13, 55, 56]. In an extreme case, A-to-I editing in filamentous fungi, the nucleotide site immediately upstream the editable A site appears to be the only *cis*-locus, where the effector allele is a T base [23, 57, 58].

For splicing-type modification, we considered a model where the unmodified isoform I_0 is converted to two modified isoforms, a functional isoform I_1 and a dysfunctional isoform I_2, at rates β_1 and β_2, respectively. As I_1 and I_2 are essentially products of the same process, their respective modification rates β_1 and β_2 are

controlled by the same *cis*-loci (Fig. 1B); thus, we assumed an allele that does not facilitate production of the I_1 will facilitate production of I_2 and vice versa. For convenience, the *cis*-genotypic value is defined as the *cis*-genotype's effect on β_1 for splicing-type modification. Hence, there is

$$\beta_1 = Q(C\hat{v} + \epsilon)$$

and

$$\beta_2 = Q(C(1 - \hat{v}) + \epsilon).$$

As splicing of a gene's transcript can be affected by a relatively large number of loci, including splicing enhancers, inhibitors, and cryptic splice sites [59, 60], we considered relatively large values of l (10, 20, 30, 40, and 50) for splicing-type modification. We assumed $\gamma_0 = 0$ but a high Q such that the I_0 only comprise a small fraction of of gene product (i.e., $P_0/(P_0 + P_1 + P_2) \approx 1\%$) to recapitulate the fact that splicing takes place co-transcriptionally [61]. We also had γ_2 significantly greater than γ_1 to reflect the effect of quality control processes, such as nonsense-mediated RNA decay [62–64], or nuclear retention and decay of intronic polyadenylated transcripts mediated by recognition of intact 5'-splice site [65–67]. The model for splicing-type modification can be readily applied as long as gene product diversity results from alternative products of an indispensable process in gene expression. For instance, it may be applied to alternative polyadenylation, in which case I_0 represents nascent RNA, and I_1 and I_2 represent RNAs polyadenylated at different sites.

Evolutionary scaling of mean modification level

When the only loci that evolve are the *cis*-loci (e.g., the *trans*-factor is invariable because of its pleiotropic effects) and the *cis*- loci's fitness effect is only mediated by gene product modification (i.e., the *cis*- loci have no pleiotropic effects), evolution of the *cis*-genotypic value v can be modeled as a discrete-state Markov process, and we can derive the probability distribution of v (and $\hat{v}$) given the initial distribution and regime of selection after evolution for a given amount of time [28, 29]. To this end, we asked what the expected relative abundance of a dysfunctional, toxic isoform (e.g., reduce fitness due to mis-interactions with other biomolecules) will be in the face of mutation, drift, and selection.

For editing-type modification, we considered a deleterious modification event that converts an unmod-

ified isoform I_0 that is functional (i.e., P_0 under stabilizing selection and fitness is a Gaussian function of $\ln P_0$; see Methods) to a modified isoform I_1 that is not functional but toxic (i.e., fitness declines with P_1; see Methods). Specifically, we examined the modification level, $f = P_1/(P_0 + P_1)$. For each combination of parameter values, we calculated the mean of v after evolution from $v = 0$ for 10^8 time steps (i.e., generations) and the corresponding f, which we refer to as a phylogenetic mean of modification level (mean modification level, for short). Under all conditions examined, the mean modification level declines with effective population size N_e (Fig. 2). Mutational bias towards the effector allele makes the mean modification level higher, whereas bias in an opposite direction makes it lower (Fig. 2A-C). For a given N_e and the per-locus mutation rate, the mean modification level becomes higher when the number of *cis*-loci, l, is high, which is most pronounced at relatively small N_e (Fig. 2A-C). This relationship between modification l is explained by the relative size of genotypic space that produce the optimal phenotype. The optimal genotype, which leads to $v = 0$, corresponds to 2^{-l} of the genotypic space. Thus, when l is large, it is harder to maintain an optimal genotype in the face of mutational pressure towards non-zero *cis*-genotypic values when l is greater [29].

Another key factor affecting the mean modification level is expression level of the gene (i.e., optimal P_0, reached when $\beta = 0$): mean modification level is lower when the gene is more highly expressed (Fig. 2D-F). This relationship is explained by toxic effect of I_1–given the modification level, there will be higher P_1 and thus greater fitness cost mediated by toxicity when the gene is highly expressed. It is a general phenomenon that high expression levels magnify the impact of errors in gene products and lead to in stronger selective constraints, as supported by previous studies of sequence evolution [68, 69], expression level [70], translation fidelity [71], as well as several types of gene product modifications [32–34].

For splicing-type modification, modification level is defined as relative abundance of the dysfunctional and toxic isoform I_2 out of all modified products, $f = P_2/(P_1 + P_2)$. As in the case of editing-type modification, mean level of splicing-type modification also declined with N_e and gene expression level, and increased with l (Fig. S1). We also compared the effect of a quality-control mechanism like nonsense-mediated decay (i.e., high γ_2) and confirmed that faster decay of I_2 can substantially lower the modification (Fig. S1). When it is the *cis*-genotypic value that is examined, results under different values of γ_2 are mostly similar (Fig. S2); when the gene's expression level is high (i.e., optimal P_1 is $\exp(4)$ or $\exp(5)$) and N_e is intermediate (i.e., 10^3-10^4), high γ_2 will have a harm-permitting effect: *cis*-genotypes that lead to production

of more I_2 will be permitted as the harmful effect is reduced by fast decay of I_2 (Fig. S2D-F, red and pink curves).

Non-monotonic scaling in *cis-trans* coevolution

Given that non-adaptive gene product diversity will be present when selection is unable to optimize the *cis*-loci in the face of mutational pressure and genetic drift, obvious questions are: why did this machinery evolve in the first place?; and how is this maintained? These are particular pertinent for editing-type modification that is not an indispensable part of gene expression. Presumably, such gene product modification processes must have additional essential functions unrelated to the set of modification events studied here, such that loss or suppression of the modification machinery will have a strongly deleterious effect. To better understand evolutionary dynamics when the modification machinery is under opposing selection forces, we considered a scenario where modification events under concern are deleterious but the *trans*-genotypic value Q is under stabilizing selection due to its contribution to an additional fitness component (Fig. 3A; also see Methods), and conducted simulations to investigate how *cis*- and *trans*-acting loci will respond to selection. We simulated evolution under different combinations of N_e, l, and strength of selection on Q. The simulation started from a high value of Q and intermediate *cis*-genotypic values (i.e., values with the largest corresponding genotypic space), representing a state that high modification activity had just evolved and optimization of *cis*-loci have not yet started.

We found the among-lineage average of Q at the end of the simulation, denoted $\bar{Q}$, is generally higher when selection on Q is strong (Fig. 3B-D, red versus blue curves). Critically, the relationship between $\bar{Q}$ and N_e is not monotonic: $\bar{Q}$ first decreases with N_e, but increases when N_e is sufficiently large. Such a relationship indicates different modes of optimization at different N_e. When N_e is too small, neither *cis*- nor *trans*- genotypic values can be efficiently optimized, so the starting condition is mostly maintained; when N_e is intermediate, as selection is still not efficient enough to optimize *cis*-loci of individual modification events in the face of mutational pressure and genetic drift, relatively low Q evolves to reduce the deleterious effect of gene product modifications globally. When N_e is sufficiently large, selection can have the population approach the global optimum where Q is optimal and modification at individual sites are optimized locally via *cis*-substitutions.

The above interpretation predicts that the tipping point where $\bar{Q}$ starts to increase with N_e should correspond to a smaller N_e when selection on Q is stronger, and that $\bar{Q}$ will be lower (given N_e) when mutational pressure is strong (i.e., l is large) and *cis*-loci are harder to optimize. Both predictions are confirmed by our simulations (Fig. 3B-D). The tipping point occurs at about $N_e = 10^{2.5}$ or $N_e = 10^3$ when selection on Q is strong (width of fitness function $\sigma_Q = 2$; see Methods), but at about $N_e = 10^4$ when selection on Q is weak ($\sigma_Q = 20$). In addition, when l is large, $\bar{Q}$ increases less with N_e after the tipping point (Fig. 3B-D).

We also examined how the deleterious modification events are shared across lineages over time. For each modification event, we calculated the fraction of lineages that shared it, and used the median across all 100 modification events to represent the level of conservation given the parameter combination (see Methods). The fraction of lineages sharing the modification generally declined over time, but declined more rapidly when N_e is large and when l is small (Fig. 3E-G, Fig. S3). When N_e is relatively small (e.g., $N_e < 10^3$) and/or l is high (e.g., $l = 10$), modifications are shared by a large proportion of, and in some case, all lineages (Fig. 3E-G, Fig. S3). Hence, when selection is too weak given mutational pressure and strength of drift, even deleterious modifications can be readily shared, and sharing of a modification event by divergent lineages may not indicate it is beneficial.

Together, our simulations of *cis-trans* coevolution demonstrate how opposing selection mediated by deleterious gene product modifications and the modification machinery's additional functions can constrain the optimization of gene product diversity. Such latent functions may explain the maintenance of some known types of gene product diversity. For instance, functional significance of protein recoding by A-to-I RNA editing has been of great interest, but most recoding events do not have any known function and are likely non-adaptive [35, 36]. A-to-I editing in non-coding RNAs transcribed from repetitive elements, however, are involved in preventing autoimmune responses [72–75] and suppressing retrotransposition [76]. In coleoids, where A-to-I editing activity in neural tissues is unusually high [21, 77], A-to-I editing is also enriched in repetitive elements [78], indicating A-to-I editing might perform similar functions in coleoids as well. Another process that is likely explainable by such a model is m6A modification, which is found to be involved in repression of endogenous retroviruses [79] and decay of mis-processed RNA [67] in a mass-action fashion. It should be noted that the above explanation does not indicate there cannot be adaptive modification events in the focal category (e.g., re-coding by A-to-I editing), as it is plausible that adaptive modifications can evolve secondarily with the modification machinery already in place. Similarly, it is compatible with

an entrenchment model [43–46] as well, as deleterious substitutions can be permitted and entrenched while the modification machinery is maintained due to its additional function; modifications that "restore" the permitted substitutions can also be considered as a latent function that contributes to the modification process's maintenance.

Our finding also revealed different optimization strategies at different N_e, which is in line with previous findings regarding global and local optimization in the evolution of quality-control mechanisms to reduce fitness cost of expression errors [80–84]. In actual biological systems, the global solution may realize as lowered expression or catalytic efficiency of the *trans*-factor, or an auto-regulatory mechanism where the *trans*-factor modifies its own gene product and trigger negative regulatory effects when its expression is too high [67, 85–87].

Simulated data recapitulate divergence of A-to-I RNA editing in coleoids

To complement our theoretical results, we asked whether simulation under our model is able to generate a distribution of modification levels that is similar to those observed in empirical studies. To this end, we examined if simulations could recapitulate the distribution of A-to-I RNA editing levels in coleoids. Previous studies reported preponderant A-to-I editing by by the ADAR family of enzymes (adenosine deaminases acting on RNA) in these coleoids' neural tissues, but the distribution of editing levels at coding sites is strongly skewed, with a vast majority of editing sites having rather low ($< 1\%$) editing levels [21, 36, 77]. We simulated evolution of 20,000 editing-type modification events, including 10,000 neutral modifications and 10,000 deleterious modifications along a phylogenetic tree of four coleoid species (Fig. 4A), with some gene-specific parameters (α, l, and C) sampled from pre-specified distributions. To reproduce a skewed distribution of modification levels like those observed in empirical studies [21, 36, 77], we sampled C from an exponential distribution with a moderate mean (i.e., magnitudes higher than ϵ but not high enough to produce an editing level above 10%). Editing levels from our simulation showed strong phylogenetic signal (i.e., neighbor-joining tree based on distance in editing levels recapitulates topology of the species tree and relative lengths of branches; Fig. 4B), and has a skewed distribution in each species (exemplified by distribution in octopus shown in Fig. 4C). Similar patterns were seen when neutral (Fig. S5A-B) and deleterious (Fig. S5C-D) editing sites were examined separately, though deleterious editing levels are generally lower. Together, our result demonstrate that observed editing level distributions can be explained by a non-adaptive model

where whole-molecule ADAR binding affinity has a skewed distribution across genes.

Concluding Remarks

In this study, we developed a theoretical model for the evolution of gene product diversity, under which we investigated how the interplay of mutations, genetic drift, and selection on isoform abundances will shape the evolutionary dynamics of gene product diversity. Our analyses of this model revealed that optimization of gene product diversity can be highly constrained by the underlying genetic architecture, the effective population size, expression level of the gene, and pleiotropic effects of the gene product modification machinery—which suggests that a substantial portion of observed gene product diversity is likely to be evolutionarily sub-optimal. Looking forward, it would be informative to conduct more comprehensive empirical analyses across a broader array of taxa; a current impediment to doing so is the lack of statistical phylogenetic tests for comparing the observed distribution of gene product diversity with that expected under the scenarios we studied. While standard statistical approaches for quantitative traits have been shown to be adequate for modeling the evolution of mRNA abundances across a phylogeny [50, 51], enabling direct comparisons between theory and data [53], this is unlikely to hold true for gene product diversity *per se*, owing to its particular genetic and mutational architecture. Our model provides a quantitative framework for developing such statistical tests.

Methods

Isoform abundances at equilibrium

Let us consider a scenario where an unmodified isoform (denoted I_0) is converted to a modified isoform (denoted I_1). Their abundances are denoted P_0 and P_1, respectively.

The rate at which P_0 changes through time is given by

$$\frac{dP_0}{dt} = \alpha - \beta P_0 - \gamma_0 P_0, \tag{1}$$

where α is the rate at which the unmodified isoform is produced, β is the net conversion rate from I_0 to I_1,

and γ_0 is the unmodified isoform's decay rate.

The rate at which P_1 changes through time is given by

$$\frac{dP_1}{dt} = \beta P_0 - \gamma_1 P_1, \tag{2}$$

where γ_1 is the modified isoform's decay rate.

An equilibrium is reached when

$$\begin{cases} \frac{dP_0}{dt} = \alpha - \beta P_0 - \gamma_0 P_0 = 0 \\ \frac{dP_1}{dt} = \beta P_0 - \gamma_1 P_1 = 0. \end{cases} \tag{3}$$

Solving the above system of equations gives

$$\begin{cases} P_0 = \frac{\alpha}{\beta+\gamma_0} \\ P_1 = \frac{\alpha\beta}{\gamma_1(\beta+\gamma_0)}. \end{cases} \tag{4}$$

The proportion of the gene product that is modified is

$$f = \frac{P_1}{P_0 + P_1} = \frac{\beta}{\beta + \gamma_1}. \tag{5}$$

The same model can be extended to more complex cases where more isoforms of the same gene's product are present. If n unique isoforms $(I_1, \ldots, I_n)$ can be produced by modifying I_0 and each molecule of I_0 can only be modified into one alternative isoform (i.e., $I_1, \ldots, I_n$ do not convert to each other), the equilibrium is reached when

$$\begin{cases} \frac{dP_0}{dt} & = \alpha - (\sum_{i=1}^{n} \beta_i) P_0 - \gamma_0 P_0 = 0 \\ \frac{dP_1}{dt} & = \beta_1 P_0 - \gamma_1 P_1 = 0 \\ & \ldots \\ \frac{dP_n}{dt} & = \beta_n P_0 - \gamma_n P_n = 0. \end{cases} \tag{6}$$

In this case, $\beta_1, \ldots, \beta_n$ are net rates at which I_0 is converted to $I_1, \ldots, I_n$, respectively, and $\gamma_1, \ldots, \gamma_n$ are decay rates of $I_1, \ldots, I_n$. The above system of equations can be rearranged and written in a matrix ($\mathbf{A}x = \mathbf{b}$)

form:

$$\begin{bmatrix} \sum_{i=1}^{n} \beta_i + \gamma_0 & 0 & \dots & 0 \\ \beta_1 & -\gamma_1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \beta_n & 0 & \dots & -\gamma_n \end{bmatrix} \begin{bmatrix} P_0 \\ P_1 \\ \dots \\ P_n \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \dots \\ 0 \end{bmatrix}. \tag{7}$$

Equilibrium abundances of different isoforms can be obtained by solving the above system of equations.

In this study, we focused on two types of gene product modification processes, editing-type and splicing-type, which are exemplified by RNA editing and splicing, respectively. A variant of the above model is applied to each of the two types. For editing-type modification, we considered a simple case with two isoforms: the unmodified isoform I_0 and the modified isoform I_1. Equilibrium isoform abundances were calculated simply using Eqn. (4). When deriving model predictions, we had $\gamma_0 = 1$ and $\gamma_1 = 1$, unless stated otherwise. For splicing-type modification, we considered a model with three isoforms: the unmodified isoform I_0 and two modified isoforms, I_1 and I_2. Equilibrium isoform abundances were calculated by solving Eqn. (7) with $n = 2$. When deriving model predictions, we had $\gamma_0 = 0$ and $\gamma_1 = 1$, unless stated otherwise.

The modeling framework also extends to multi-step modification, where an modified isoform can be further modified into a different one. Let us consider a scenario where a modified isoform I_1 is modified into an different isoform I_2. The equilibrium is reached when

$$\begin{cases} \frac{dP_0}{dt} = \alpha - \beta_{0\to1} P_0 - \gamma_0 P_0 = 0 \\ \frac{dP_1}{dt} = \beta_{0\to1} P_0 - \beta_{1\to2} P_1 - \gamma_1 P_1 = 0 \\ \frac{dP_2}{dt} = \beta_{1\to2} P_1 - \gamma_2 P_2 = 0. \end{cases} \tag{8}$$

Solving the above system of equations gives

$$\begin{cases} P_0 = \frac{\alpha}{\beta_{0\to1}+\gamma_0} \\ P_1 = \frac{\alpha\beta_{0\to1}}{(\beta_{0\to1}+\gamma_0)(\beta_{1\to2}+\gamma_1)} \\ P_2 = \frac{\alpha\beta_{0\to1}\beta_{1\to2}}{(\beta_{0\to1}+\gamma_0)(\beta_{1\to2}+\gamma_1)\gamma_2}. \end{cases} \tag{9}$$

Similarly, if there is a series of n modified isoforms where I_i is produced by modifying I_{i-1}:

$$\begin{cases} \frac{dP_0}{dt} & = \alpha - \beta_{0\to1}P_0 - \gamma_0 P_0 = 0 \\ \frac{dP_1}{dt} & = \beta_{0\to1}P_0 - \beta_{1\to2}P_1 - \gamma_1 P_1 = 0 \\ & \dots \\ \frac{dP_{n-1}}{dt} & = \beta_{n-2\to n-1}P_{n-2} - \beta_{n-1\to n}P_{n-1} - \gamma_{n-1}P_{n-1} = 0 \\ \frac{dP_n}{dt} & = \beta_{n-1\to n}P_{n-1} - \gamma_n P_n = 0. \end{cases} \tag{10}$$

The above system of equations can be rearranged and written in a matrix ($\mathbf{A}x = \mathbf{b}$) form:

$$\begin{bmatrix} \beta_{0\to1}+\gamma_0 & 0 & \dots & 0 & 0 & 0 \\ \beta_{0\to1} & -\beta_{1\to2}-\gamma_1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \beta_{n-2\to n-1} & -\beta_{n-1\to n}-\gamma_{n-1} & 0 \\ 0 & 0 & \dots & 0 & \beta_{n-1\to n} & -\gamma_n \end{bmatrix} \begin{bmatrix} P_0 \\ P_1 \\ \dots \\ P_{n-1} \\ P_n \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix}. \tag{11}$$

It is worth noting that the above model can be applied when it is the number of modification events within the same RNA or protein molecule but not the exact locations of the modifications that is of interest. In such a case, n represents the total number of sites in the RNA or protein molecule that can potentially be modified, and I_i represents isoforms where i of the n potential sites are modified. If the per-site modification rate is constant regardless of location of the potential modification site or modification states of other sites, such that for each $0 \le i \le n-1$ there is $\beta_{i\to i+1} = (n-i)\beta$, Eqn. (10) and (11) can be written as

$$\begin{cases} \frac{dP_0}{dt} & = \alpha - n\beta P_0 - \gamma_0 P_0 = 0 \\ \frac{dP_1}{dt} & = n\beta P_0 - (n-1)\beta P_1 - \gamma_1 P_1 = 0 \\ & \dots \\ \frac{dP_{n-1}}{dt} & = 2\beta P_{n-2} - \beta P_{n-1} - \gamma_{n-1}P_{n-1} = 0 \\ \frac{dP_n}{dt} & = \beta P_{n-1} - \gamma_n P_n = 0 \end{cases} \tag{12}$$

and

$$\begin{bmatrix} n\beta+\gamma_0 & 0 & \dots & 0 & 0 & 0 \\ n\beta & -(n-1)\beta-\gamma_1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 2\beta & -\beta-\gamma_{n-1} & 0 \\ 0 & 0 & \dots & 0 & \beta & -\gamma_n \end{bmatrix} \begin{bmatrix} P_0 \\ P_1 \\ \dots \\ P_{n-1} \\ P_n \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix}. \tag{13}$$

In the most general form of the model where every isoform (e.g., I_i) can be converted to another isoform (e.g., I_j) at per-molecule rate $\beta_{i,j}$ ($\beta_{i,j} = 0$ if $i = j$), Eqn. (7) will be written as

$$\begin{bmatrix} \sum_{i=1}^{n}\beta_{0,i}+\gamma_0 & 0 & \dots & 0 \\ \beta_{0,1} & -\sum_{i=0}^{n}\beta_{1,i}-\gamma_1 & \dots & \beta_{n,1} \\ \dots & \dots & \dots & \dots \\ \beta_{0,n} & \beta_{1,n} & \dots & -\sum_{i=0}^{n}\beta_{n,i}-\gamma_n \end{bmatrix} \begin{bmatrix} P_0 \\ P_1 \\ \dots \\ P_n \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \dots \\ 0 \end{bmatrix}. \tag{14}$$

Genetic architecture of modification rate

For a given modified isoform, the corresponding β parameter is determined together by l *cis*-acting loci and a *trans*-genotypic value, Q. The *trans*-genotypic value Q characterizes the overall activity of the enzyme (or molecular machinery) that carries out the modification process, and is a product of its expression level and per-molecule activity (i.e., the catalysis efficiency of an enzyme, determined by its amino acid sequence and/or conformation). The binding affinity between the enzyme and its substrate is dependent on the *cis*-loci, which are sites in regions adjacent to (though not necessarily immediately adjacent to) the site subject to modification.

We assumed that each *cis*-locus can have either an "effector" allele that facilitates binding between the modification enzyme and its substrate (i.e., sequence motif recognized by the enzyme), or a "null" allele that does not facilitate binding. The total effect of the *cis*-loci on β is determined by a normalized genotypic value $\hat{v}$, which is calculated as

$$\hat{v} = \frac{v}{v_{max}}, \tag{15}$$

where v is the sum of all effector alleles' effect, and v_{max} is the greatest possible value of v (i.e., achieved

when there are no null alleles). In this study, we focused on a simple model where the *cis*-loci's effect is additive and all *cis*-loci have equal effect, so there v is equal to the total number of effector alleles and v_{max} is equal to the number of *cis*-loci, l.

The relationship between β and underlying parameters is given by

$$\beta = Q(C\hat{v} + \epsilon). \tag{16}$$

Here, $\epsilon \geq 0$ is the rate of non-specific modification that takes place independent of the *cis*-genotype, and $C > 0$ reflects global structural features of an RNA or protein molecule that affect binding affinity between the enzyme and the substrate.

For splicing-type modification, we assumed that β_1 and β_2 are affected by the same set of *cis*-loci. We also assumed the two alleles that each *cis*-locus could potentially have are both effector alleles. One of them only facilitates the production of I_1, whereas the other only facilitates the production of I_2; under this model, the same genotype's effects on I_1 and I_2 are inversely correlated. For convenience, we defined the normalized *cis*-genotypic value based on the genotype's effect on β_1. The β parameters are thus given by

$$\beta_1 = Q(C\hat{v} + \epsilon) \tag{17}$$

and

$$\beta_2 = Q(C(1 - \hat{v}) + \epsilon). \tag{18}$$

For editing-type modification, we had $C = 1$, $Q = 1$, and $epsilon = 0$ when deriving model predictions, unless specified otherwise. For splicing-type, we had $C = 1$, $Q = 100$, and $epsilon = 0$, unless specified otherwise.

The mutational spectrum of each *cis*-locus is characterized by two per-locus mutation rates: μ_{01}, the rate of mutations from the null allele to the effector allele, and μ_{10}, the rate of mutations in the opposite direction. The difference between μ_{01} and μ_{10} reflects difference in the two allele's corresponding sequence spaces (i.e., number of nucleotide states, assuming each locus represents a single site) and/or rate of different types of nucleotide changes (e.g., transition/transversion bias or AT-bias). In the case of splicing-type modification, μ_{01} and μ_{10} are simply replaced by mutation rates between two effector alleles. For simplicity,

we assumed that all *cis*-loci have the same mutational spectrum in this study. In this study, we had the total mutation rate per locus $\mu_{01} + \mu_{10} = 2 \times 10^{-9}$, unless stated otherwise.

Selection on isoform abundance

We first considered a scenario where each isoform contributes to fitness independently, in which case the fitness is given by

$$\omega = \prod_{i=0}^{n} \omega_i, \tag{19}$$

where ω_i is fitness with respect to P_i.

We considered two scenarios where an isoform's abundance is subject to selection: a scenario where the isoform is functional and a scenario where it is not functional but deleterious. Selection on the abundance of a functional isoform I_i is characterized by a Gaussian fitness function:

$$\omega_i = \exp\left(-\frac{\ln P_i - \ln P_{i.opt}}{2\sigma_i^2}\right), \tag{20}$$

where σ_i^2, the fitness function's width and characterizes the strength of selection.

If I_i is deleterious, fitness with respect to its abundance P_i is given by

$$\omega_i = \exp(-\lambda_i P_i), \tag{21}$$

where $\lambda_i > 0$ is a parameter characterizing the strength of selection. When $\lambda_i = 0$, there is $\omega_i = 1$, which corresponds to the case that P_i is not under selection. In this study, we had $\sigma = 10$ for every functional isoform and $\lambda = 10^{-3}$ for every deleterious isoform, unless specified otherwise.

For editing-type modifications, we mainly focused on a scenario where the modification is deleterious– i.e., I_0 is functional while I_1 is toxic. Fitness component with respect to P_0 is calculated by Eqn. 20), whereas fitness with respect to P_1 is calculated by Eqn. 21). For splicing-type modifications, fitness is determined only by P_1 and P_2, not P_0. One of the modified isoforms, I_1, is the functional, and its abundance P_1 is under stabilizing selection; fitness component with respect to P_1 is thus computed using Eqn. 20). The other modified isoform, I_2, in contrast, is not functional but toxic, and the corresponding fitness component

is computed using Eqn. 21). With I_2 representing mis-processed isoform(s), we also assumed that γ_2 is greater than γ_1 to recapitulate quality-control mechanisms that act to eliminate mis-processed isoforms [62–64]; specifically, we examined scenarios of $\gamma_1 = 1$ while γ_2 is equal to 20, 50, or 100.

Distribution of *cis*-genotypic value

When the number of *cis*-loci underlying a modification event is reasonably small, the evolution of genotypic value v (and thus $\hat{v}$) can be approximated by a sequential-fixation (strong-selection-weak-mutation) model [88]. Then, assuming that other parameters that affect modification are constant, the evolution of v (and $\hat{v}$) can be modeled as a Markov process with a constant transition matrix. A time step in this Markov process can be a generation or any arbitrary time interval as long as the the probability that more than one mutations arise in the population is very low (i.e., $2N_e\mu < 0.01$) such that the sequential-fixation model is an appropriate approximation [29]. Using this approach, the distribution of the *cis*-genotypic value v given the starting state after a given amount of time can be derived.

Let us consider a simple scenario where the effector allele at every *cis*-locus has an effect size of 1 (i.e., v is equal to the number of effector alleles and $v_{max} = l$). In a diploid population, the probability that v becomes $v + 1$ via a substitution in a time step given the present genotypic value v is

$$\Pr(v \to v+1) = 2(l-v)N_e u_{01} f_{v \to v+1}, \tag{22}$$

where N_e is the effective population size and $f_{v \to v+1}$ is the fixation probability given ancestral and mutant phenotypes.

Similarly, the probability of becoming $v - 1$ via a substitution is

$$\Pr(v \to v-1) = 2vN_e u_{10} f_{v \to v-1}. \tag{23}$$

The probability that v does not change is simply

$$\Pr(v \to v) = 1 - \Pr(v \to v+1) - \Pr(v \to v-1). \tag{24}$$

The fixation probability is obtained using Kimura's method [89]:

$$\Pr(\text{fixation}|N_e, s) = \frac{1 - \exp(-2s)}{1 - \exp(-4N_e s)},$$

where $s = \frac{\omega_M}{\omega_A} - 1$ is the coefficient of selection (ω_M and ω_A represent mutant and ancestral fitness, respectively).

Given the probability distribution of v at a time t, $\mathbf{v}_t$, the distribution at $t+1$ is

$$\mathbf{v}_{t+1} = \mathbf{v}_t \mathbf{T}, \tag{25}$$

where $\mathbf{v}_t$ and $\mathbf{v}_{t+1}$ are row vectors of length $l+1$, with each element represents the probability of a possible value of v. The transition matrix $\mathbf{T}$ is a $l+1 \times l+1$ matrix where $\mathbf{T}[i+1, j+1] = \Pr(i \to j)$. The probability $\Pr(i \to j)$ is calculated following Eqn. 22 and 23 if $0 \leq i \leq l$, $0 \leq j \leq l$, and $|i-j| \leq 1$; otherwise, $\Pr(i \to j) = 0$. In this study, we used $\mathbf{v}_0 \mathbf{T}^{1e8}$ to represent an equilibrium distribution. For editing-type modification, we had the first element of $\mathbf{v}_0$ equal to 1 (i.e., starting from the genotype that has the least effect on modification), whereas for splicing-type modification, we had the last element of $\mathbf{v}_0$ equal to 1 (i.e., starting from the genotype that maximizes the production of I_1 and minimized the production of I_2).

If different *cis*-loci have different effect sizes, there will be up to $\binom{l}{2}$ possible values of v. In the extreme case where all loci have different effect sizes and the mutation rate varies depending both on the locus and the ancestral allele, the transition probability from a given genotype to a given neighbor genotype (i.e., differ from the ancestral genotype at only one site) is simply the product of the mutation rate (given the locus and the ancestral allele) and the fixation probability. In this study, we focused mostly on the simple scenario where all the *cis*-loci have equal effect size and mutation rates, although the modeling framework can be readily extended to more general cases.

Simulating *cis-trans* coevolution

To investigate coevolutionary dynamics between the *cis*-loci and the *trans*-genotypic value Q when many genes or sites are subject to modification, we conducted simulations of evolution where *cis*-loci and Q are

both affected by mutations.

Each lineage we simulated was divided into a number of time steps, with the number of time steps proportional to the branch length. If the only loci that could undergo evolutionary changes in a time interval are the *cis*-loci, the probability distribution of a given modification event's *cis*-genotypic value v at the end of the time interval is simply

$$\mathbf{v}_t = \mathbf{v}_0 \mathbf{T}^t \tag{26}$$

where v_0 is the starting distribution and t is the number of time steps the time interval consists of. If the simulation starts from a pre-designated value of v, the corresponding element of $\mathbf{v}_0$ will be 1 while other elements are equal to 0.

Before simulating evolution for a lineage, we first determined m, the total number of mutations that affect Q to occur during evolution by sampling m from a Poisson distribution with mean equal to $2N_e U_Q L$, where L is the branch length (i.e., number of time steps) and U_Q is the rate of mutations that affect Q. Then we randomly picked m time steps, at each of which a mutation affecting Q would occur. If $m > L$ (which has very low probability given parameter values considered, and did not happen in our simulations), this value of m will not be used for simulations. The effect of each mutation on $\ln Q$ was sampled from a normal distribution $\mathcal{N}(0, \mathrm{S}_Q)$. Change in the distribution v during the interval between two mutations that affect Q obtained using Eqn. 26, with t being the number of time steps between two mutations. Before examining the fitness effect of a mutation that affects Q, a value of v was first sampled from its distribution, which, together with the mutation's effect on Q, will determine the fixation probability. If the mutation is fixed, the transition matrix will be re-calculated with the mutant Q, and the mutant Q will be the new Q to begin with when the next mutation is examined. When products of multiple genes are subject to modification, fitness effect of each mutation affecting Q is determined collectively by its effect on all modification events; when such a mutation is fixed, all gene's transition matrices will be altered. For simplicity, we assumed that different modification events' *cis*-loci are not shared and evolve independently (i.e., no linkage between *cis*-mutations affecting different modification events).

We considered a scenario where the modification machinery has both beneficial and detrimental effects on fitness at the same time. Under this model, there are a set of genes subject to deleterious editing-type modifications (i.e., the unmodified isoform is functional and the modified isoform is deleterious). At the

same time, Q contributes to a fitness component ω_Q that is independent of these modification events. In our simulations, Q was under stabilizing selection, and ω_Q is given by

$$\omega_Q = \exp\left(-\frac{\ln Q - \ln Q_{opt}}{2\sigma_Q^2}\right), \tag{27}$$

where Q_{opt} is the optimal value of Q and σ_Q is the fitness function's width. In this case, if there are n genes subject to modification, the overall fitness is given by

$$\omega = \omega_Q \prod_{i=0}^{n} \omega_i, \tag{28}$$

where ω_i is fitness with respect to the i-th gene's isoform abundances.

Values of N_e used in the simulations include 10^2, $10^{2.5}$, 10^3, $10^{3.5}$, 10^4, $10^{4.5}$, and 10^5. In each simulation, we considered 100 genes that are subject to deleterious modifications. For simplicity, we had all modification events have equal l, and considered scenarios of $l = 2$, $l = 5$, and $l = 10$, where the initial value of v was 1, 2, and 5, respectively. Regarding selection on Q, we considered two scenarios: a scenario of strong selection ($\sigma_Q = 2$) and a scenario of relatively weak selection ($\sigma_Q = 20$). In all simulations, we had $Q_{opt} = 2$, $U_Q = 10^{-8}$ and $\mathrm{S}_Q = 0.1$. We also had $\alpha = 1$, $\gamma_0 = 1$, $\gamma_1 = 1$, $C = 1$, $\sigma = 10$, $\lambda = 10^{-3}$, and $\epsilon = 10^{-3}$ for all genes in all simulations. Starting value of Q was equal to its optimum for all simulations. After the simulations, we quantified the degree to which the modifications are shared (i.e., conserved) among lineages. For each gene, we calculated the the fraction of lineages where $P_1 > 0.005$. The median of all genes is then used to represent how likely a modification event is shared given the evolutionary parameters (l, N_e, and strength of selection). We examined how this value varied depending on divergence time by performing the simulation with different times of duration, including 2×10^7, 4×10^7, 6×10^7, 8×10^7, and 10^8 time steps. For each combination of parameter values, we simulated 50 independent lineages.

The above procedure can also be used to simulate the coevolution of the *cis*-loci and other parameters, such as α, C or ϵ, in which case mutations affecting Q in the above procedure will be replaced by mutations affecting the parameter of interest.

Simulation along the coleoid tree

We simulated evolution of editing levels at $20,000$ editing sites along a phylogenetic tree of four coleoid species: the common octopus (*Octopus vulgaris*), the bimac (*O. bimaculoides*), the squid (*Doryteuthis pealeii*), and the cuttlefish (*Sepia oficianalis*). The coleoids have high A-to-I RNA editing activity in their neural tissues, whereas extant non-coleoid cephalopods and non-cephalopod mollusks do not [21, 77]. Branch lengths of the phylogenetic tree are based on divergence times described in ref. [21], with mid point the reported range used for our simulations. Divergence time of the octopus and the bimac, which are very closely related, was set to be 5 million years. We assumed each time step in the simulation corresponds to a year, so the number of time steps a branch corresponds to is equal to branch length in terms of years. We started the simulation from the most recent common ancestor of four coleoids, and the value of v of each editing site at this ancestral node was sampled randomly from the corresponding genotypic space. We assumed that Q is under strong stabilizing selection mediated by functions independent of the focal editing events such that Q remained constant in the simulation. We had $Q = 1$ for this simulation. The distribution of v at the end of each branch was obtained using Eqn. (26) with time of evolution equal to branch length; a value of v was then sampled from the distribution to represent the state at the end of this branch and the starting state of its descendent branches (if any). Some gene-specific parameters were sampled from pre-specified distributions. The rate at which I_0 is expressed, α, was sampled from a log-normal distribution; that is, $\ln \alpha$ was sampled from $\mathcal{N}(0, 1)$. The number of *cis*-loci, l, was sampled uniformly from $(0, 1, \ldots, 10)$. The C parameter was sampled from a exponential distribution with mean equal to 0.1. All genes had $\gamma_0 = 1$, $\gamma_1 = 1$, $\sigma = 10$, $\lambda = 10^{-3}$, and $\epsilon = 10^{-4}$. Because $\epsilon > 0$, all editing levels were positive. Thus, after the simulation, we log-transformed all editing levels and computed Euclidean distances between each pair of species using log-transformed editing levels ($\ln(f)$). We then built a neighbor-joining (NJ) tree based on these distances using the *nj* function of R package *ape*, and asked this NJ tree recapitulate the phylogenetic relationship of the four coleoid species; specifically, we examined whether 1) the two *Octopus* species fall in one clade while the squid and the cuttlefish fall in another, and 2) whether distance between the two octopuses is closer than that between the squid and the cuttlefish.

Code and data availability

Code and data files are available at https://github.com/phylo-lab-usc/gene_product_diversity

Acknowledgements

We thank Mark Kim, and members of the Pennell, Edge, and Mooney labs for their thoughtful comments on parts of this study. We acknowledge support from the Natural Sciences and Engineering Research Council of Canada (FN 492860, to AFP), the Jean D'Alembert Foundation (France 2030 program ANR-11-IDEX-0003, to AFP), and the National Institute of General Medical Sciences (R35GM151348, to MP).

References

[1] Zhang, J & Xu, C. (2022) *Trends in Genetics*.

[2] The FANTOM Consortium and the RIKEN PMI and CLST (DGT). (2014) *Nature* **507**, 462–470.

[3] Kimura, K, Wakamatsu, A, Suzuki, Y, Ota, T, Nishikawa, T, Yamashita, R, Yamamoto, J.-i, Sekine, M, Tsuritani, K, Wakaguri, H, et al. (2006) *Genome research* **16**, 55–65.

[4] Scotti, M. M & Swanson, M. S. (2016) *Nature Reviews Genetics* **17**, 19–32.

[5] Kalsotra, A & Cooper, T. A. (2011) *Nature Reviews Genetics* **12**, 715–729.

[6] Barbosa-Morais, N. L, Irimia, M, Pan, Q, Xiong, H. Y, Gueroussov, S, Lee, L. J, Slobodeniuc, V, Kutter, C, Watt, S, Colak, R, et al. (2012) *Science* **338**, 1587–1593.

[7] Wright, C. J, Smith, C. W, & Jiggins, C. D. (2022) *Nature Reviews Genetics* **23**, 697–710.

[8] Goldtzvik, Y, Sen, N, Lam, S. D, & Orengo, C. (2023) *Current Opinion in Structural Biology* **81**, 102640.

[9] Di Giammartino, D. C, Nishida, K, & Manley, J. L. (2011) *Molecular cell* **43**, 853–866.

[10] Farajollahi, S & Maas, S. (2010) *Trends in Genetics* **26**, 221–230.

[11] Nishikura, K. (2010) *Annual review of biochemistry* **79**, 321–349.

[12] Nishikura, K. (2016) *Nature reviews Molecular cell biology* **17**, 83–96.

[13] Li, S & Mason, C. E. (2014) *Annual review of genomics and human genetics* **15**, 127–150.

[14] Lee, S, Liu, B, Lee, S, Huang, S.-X, Shen, B, & Qian, S.-B. (2012) *Proceedings of the National Academy of Sciences* **109**, E2424–E2432.

[15] Mann, M & Jensen, O. N. (2003) *Nature biotechnology* **21**, 255–261.

[16] Gout, J.-F, Li, W, Fritsch, C, Li, A, Haroon, S, Singh, L, Hua, D, Fazelinia, H, Smith, Z, Seeholzer, S, et al. (2017) *Science advances* **3**, e1701484.

[17] Dunn, J. G, Foo, C. K, Belletier, N. G, Gavis, E. R, & Weissman, J. S. (2013) *elife* **2**, e01179.

[18] Allan Drummond, D & Wilke, C. O. (2009) *Nature Reviews Genetics* **10**, 715–724.

[19] de Pouplana, L. R, Santos, M. A, Zhu, J.-H, Farabaugh, P. J, & Javid, B. (2014) *Trends in biochemical sciences* **39**, 355–362.

[20] de Klerk, E & AC't Hoen, P. (2015) *Trends in Genetics* **31**, 128–139.

[21] Liscovitch-Brauer, N, Alon, S, Porath, H. T, Elstein, B, Unger, R, Ziv, T, Admon, A, Levanon, E. Y, Rosenthal, J. J, & Eisenberg, E. (2017) *Cell* **169**, 191–202.

[22] Garrett, S & Rosenthal, J. J. (2012) *Science* **335**, 848–851.

[23] Liu, H, Li, Y, Chen, D, Qi, Z, Wang, Q, Wang, J, Jiang, C, & Xu, J.-R. (2017) *Proceedings of the National Academy of Sciences* **114**, E7756–E7765.

[24] Xin, K, Zhang, Y, Fan, L, Qi, Z, Feng, C, Wang, Q, Jiang, C, Xu, J.-R, & Liu, H. (2023) *Proceedings of the National Academy of Sciences* **120**, e2219029120.

[25] Salz, H. K. (2011) *Current opinion in genetics & development* **21**, 395–400.

[26] Hansen, T. B, Jensen, T. I, Clausen, B. H, Bramsen, J. B, Finsen, B, Damgaard, C. K, & Kjems, J. (2013) *Nature* **495**, 384–388.

[27] Kristensen, L. S, Andersen, M. S, Stagsted, L. V, Ebbesen, K. K, Hansen, T. B, & Kjems, J. (2019) *Nature reviews genetics* **20**, 675–691.

[28] Lynch, M & Hagner, K. (2015) *Proceedings of the National Academy of Sciences* **112**, E30–E38.

[29] Lynch, M. (2020) *Proceedings of the National Academy of Sciences* **117**, 10435–10444.

[30] Saudemont, B, Popa, A, Parmley, J. L, Rocher, V, Blugeon, C, Necsulea, A, Meyer, E, & Duret, L. (2017) *Genome biology* **18**, 1–15.

[31] Xu, C & Zhang, J. (2021) *Cell reports* **36**.

[32] Bénitière, F, Necsulea, A, & Duret, L. (2024) *Elife* **13**, RP93629.

[33] Pickrell, J. K, Pai, A. A, Gilad, Y, & Pritchard, J. K. (2010) *PLoS genetics* **6**, e1001236.

[34] Xu, C & Zhang, J. (2018) *Cell systems* **6**, 734–742.

[35] Xu, G & Zhang, J. (2014) *Proceedings of the National Academy of Sciences* **111**, 3769–3774.

[36] Jiang, D & Zhang, J. (2019) *Nature Communications* **10**, 5411.

[37] Nguyen, T. A, Heng, J. W. J, Ng, Y. T, Sun, R, Fisher, S, Oguz, G, Kaewsapsak, P, Xue, S, Reversade, B, Ramasamy, A, et al. (2023) *BMC biology* **21**, 251.

[38] Liu, Z & Zhang, J. (2018) *Molecular Biology and Evolution* **35**, 963–969.

[39] Liu, J, Dou, X, Chen, C, Chen, C, Liu, C, Xu, M. M, Zhao, S, Shen, B, Gao, Y, Han, D, et al. (2020) *Science* **367**, 580–586.

[40] Wang, X, Lu, Z, Gomez, A, Hon, G. C, Yue, Y, Han, D, Fu, Y, Parisien, M, Dai, Q, Jia, G, et al. (2014) *Nature* **505**, 117–120.

[41] Landry, C. R, Levy, E. D, & Michnick, S. W. (2009) *Trends in genetics* **25**, 193–197.

[42] Landry, C. R, Freschi, L, Zarin, T, & Moses, A. M. (2014) *Frontiers in genetics* **5**, 104097.

[43] Stoltzfus, A. (1999) *Journal of molecular evolution* **49**, 169–181.

[44] Lukeš, J, Archibald, J. M, Keeling, P. J, Doolittle, W. F, & Gray, M. W. (2011) *IUBMB life* **63**, 528–537.

[45] Wideman, J. G, Novick, A, Muñoz-Gómez, S. A, & Doolittle, W. F. (2019) *Current Opinion in Genetics & Development* **58**, 87–94.

[46] Muñoz-Gómez, S. A, Bilolikar, G, Wideman, J. G, & Geiler-Samerotte, K. (2021) *Journal of molecular evolution* **89**, 172–182.

[47] Covello, P. S & Gray, M. W. (1989) *Nature* **341**, 662–666.

[48] Fiebig, A, Stegemann, S, & Bock, R. (2004) *Nucleic acids research* **32**, 3615–3622.

[49] Gray, M. W. (2012) *Biochemistry* **51**, 5235–5242.

[50] Chen, J, Swofford, R, Johnson, J, Cummings, B. B, Rogel, N, Lindblad-Toh, K, Haerty, W, Di Palma, F, & Regev, A. (2019) *Genome research* **29**, 53–63.

[51] Dimayacyac, J. R, Wu, S, Jiang, D, & Pennell, M. (2023) *Genome Biology and Evolution* **15**, evad211.

[52] Jiang, D, Cope, A. L, Zhang, J, & Pennell, M. (2023) *Molecular Biology and Evolution* **40**, msad169.

[53] Price, P. D, Palmer Droguett, D. H, Taylor, J. A, Kim, D. W, Place, E. S, Rogers, T. F, Mank, J. E, Cooney, C. R, & Wright, A. E. (2022) *Nature Ecology & Evolution* **6**, 1035–1045.

[54] Cope, A. L, Schraiber, J. G, & Pennell, M. (2024) *bioRxiv* p. doi:10.1101/2024.07.08.602411.

[55] Lehmann, K. A & Bass, B. L. (2000) *Biochemistry* **39**, 12875–12884.

[56] Linder, B, Grozhik, A. V, Olarerin-George, A. O, Meydan, C, Mason, C. E, & Jaffrey, S. R. (2015) *Nature methods* **12**, 767–772.

[57] Wang, C, Xu, J.-R, & Liu, H. (2016) *RNA biology* **13**, 940–945.

[58] Liu, H, Wang, Q, He, Y, Chen, L, Hao, C, Jiang, C, Li, Y, Dai, Y, Kang, Z, & Xu, J.-R. (2016) *Genome research* **26**, 499–509.

[59] Wang, J, Smith, P. J, Krainer, A. R, & Zhang, M. Q. (2005) *Nucleic acids research* **33**, 5053–5062.

[60] Wang, Z & Burge, C. B. (2008) *Rna* **14**, 802–813.

[61] Herzel, L, Ottoz, D. S, Alpert, T, & Neugebauer, K. M. (2017) *Nature reviews Molecular cell biology* **18**, 637–650.

[62] Frischmeyer, P. A & Dietz, H. C. (1999) *Human molecular genetics* **8**, 1893–1900.

[63] Kurosaki, T & Maquat, L. E. (2016) *Journal of cell science* **129**, 461–467.

[64] Kurosaki, T, Popp, M. W, & Maquat, L. E. (2019) *Nature reviews Molecular cell biology* **20**, 406–420.

[65] Lee, E. S, Akef, A, Mahadevan, K, & Palazzo, A. F. (2015) *PLoS One* **10**, e0122743.

[66] Lee, E. S, Smith, H. W, Wolf, E. J, Guvenek, A, Wang, Y. E, Emili, A, Tian, B, & Palazzo, A. F. (2022) *RNA* **28**, 878–894.

[67] Lee, E. S, Smith, H. W, Ihn, S. S, Scalize de Olivera, L, Wang, Y. E, Jomphe, R. Y, Nabeel-Shah, S, Pu, S, Greenblatt, J. F, & Palazzo, A. F. (2023) *bioRxiv* pp. 2023–06.

[68] Zhang, J & Yang, J.-R. (2015) *Nature Reviews Genetics* **16**, 409–420.

[69] Managadze, D, Rogozin, I. B, Chernikova, D, Shabalina, S. A, & Koonin, E. V. (2011) *Genome biology and evolution* **3**, 1390–1404.

[70] Liao, B.-Y & Zhang, J. (2006) *Molecular biology and evolution* **23**, 1119–1128.

[71] Mordret, E, Dahan, O, Asraf, O, Rak, R, Yehonadav, A, Barnabas, G. D, Cox, J, Geiger, T, Lindner, A. B, & Pilpel, Y. (2019) *Molecular Cell* **75**, 427–441.

[72] de Reuver, R, Verdonck, S, Dierick, E, Nemegeer, J, Hessmann, E, Ahmad, S, Jans, M, Blancke, G, Van Nieuwerburgh, F, Botzki, A, et al. (2022) *Nature* **607**, 784–789.

[73] Karki, R, Sundaram, B, Sharma, B. R, Lee, S, Malireddi, R. S, Nguyen, L. N, Christgen, S, Zheng, M, Wang, Y, Samir, P, et al. (2021) *Cell reports* **37**.

[74] Liddicoat, B. J, Piskol, R, Chalk, A. M, Ramaswami, G, Higuchi, M, Hartner, J. C, Li, J. B, Seeburg, P. H, & Walkley, C. R. (2015) *Science* **349**, 1115–1120.

[75] Chung, H, Calis, J. J, Wu, X, Sun, T, Yu, Y, Sarbanes, S. L, Thi, V. L. D, Shilvock, A. R, Hoffmann, H.-H, Rosenberg, B. R, et al. (2018) *Cell* **172**, 811–824.

[76] Orecchini, E, Frassinelli, L, & Michienzi, A. (2017) *RNA biology* **14**, 1485–1491.

590 [77] Alon, S, Garrett, S. C, Levanon, E. Y, Olson, S, Graveley, B. R, Rosenthal, J. J, & Eisenberg, E. (2015)
591 *Elife* **4**, e05198.

592 [78] Albertin, C. B, Medina-Ruiz, S, Mitros, T, Schmidbaur, H, Sanchez, G, Wang, Z. Y, Grimwood, J,
593 Rosenthal, J. J, Ragsdale, C. W, Simakov, O, et al. (2022) *Nature communications* **13**, 2427.

594 [79] Chelmicki, T, Roger, E, Teissandier, A, Dura, M, Bonneville, L, Rucli, S, Dossin, F, Fouassier, C,
595 Lameiras, S, & Bourc'his, D. (2021) *Nature* **591**, 312–316.

596 [80] Rajon, E & Masel, J. (2011) *Proceedings of the National Academy of Sciences* **108**, 1082–1087.

597 [81] Xiong, K, McEntee, J. P, Porfirio, D. J, & Masel, J. (2017) *Genetics* **205**, 397–407.

598 [82] Ho, A. T & Hurst, L. D. (2021) *Molecular Biology and Evolution* **38**, 244–262.

599 [83] Koonin, E. V. (2006) *Biology direct* **1**, 1–23.

600 [84] Koonin, E. V. (2016) *BMC biology* **14**, 1–8.

601 [85] Ni, J. Z, Grate, L, Donohue, J. P, Preston, C, Nobida, N, O'Brien, G, Shiue, L, Clark, T. A, Blume, J. E,
602 & Ares, M. (2007) *Genes & development* **21**, 708–718.

603 [86] Carvill, G. L & Mefford, H. C. (2020) *Current opinion in genetics & development* **65**, 98.

604 [87] Lareau, L. F, Inada, M, Green, R. E, Wengrod, J. C, & Brenner, S. E. (2007) *Nature* **446**, 926–929.

605 [88] McCandlish, D. M & Stoltzfus, A. (2014) *The Quarterly review of biology* **89**, 225–252.

606 [89] Kimura, M. (1962) *Genetics* **47**, 713.

Tables

Table 1: Definitions and notations of parameters.

Parameter	Definition
I_i	The i-th modified isoform; I_0 represents the unmodified isoform.
P_i	Abundance of I_i.
α	Rate at which P_0 is produced.
β_i	Per-molecule net rate at which the I_0 is converted to the I_i.
γ_i	Decay rate of I_i.
f	Modification level; $f = \frac{P_1}{P_1+P_0}$ for editing-type and $f = \frac{P_2}{P_1+P_2}$ for splicing-type.
l	Number of *cis*-loci affecting β.
v	*cis*-genotypic value characterizing the combined effect of the *cis*-genotype on β.
v_{max}	Value of v when every locus has an effector allele.
$\hat{v}$	Normalized *cis*-genotypic value, $\hat{v} = \frac{v}{v_{max}}$.
Q	*trans*-genotypic value underlying β.
C	Parameter characterizing gene-level feature that affect *cis*-loci's effect size on β.
μ_{01}	Mutation rate from null allele to effector allele per *cis*-loci.
μ_{10}	Mutation rate from null allele to effector allele per *cis*-loci.
ω	Overall fitness.
ω_i	Fitness with respect to P_i.
σ_i	Width of the fitness function when P_i is under stabilizing selection.
λ_i	Parameter characterizing speed at which ω_i declines with P_i when I_i is toxic.
s	Coefficient of selection of a mutation.
N_e	Effective population size.
$\Pr(i \to j)$	Probability that v changes from i to j via a substitution in a time step.
$\mathbf{T}$	Transition matrix for the genotypic value v.
$\mathbf{v}_t$	Probability distribution of v at a given time t.
v_t	Value of v at a given time t; v_0 is the starting value.
U_Q	Rate of mutations affecting Q.
S_Q	Standard deviation of mutation's effect on $\ln Q$.
σ_Q	Width of the fitness function of Q.

608 # Figures

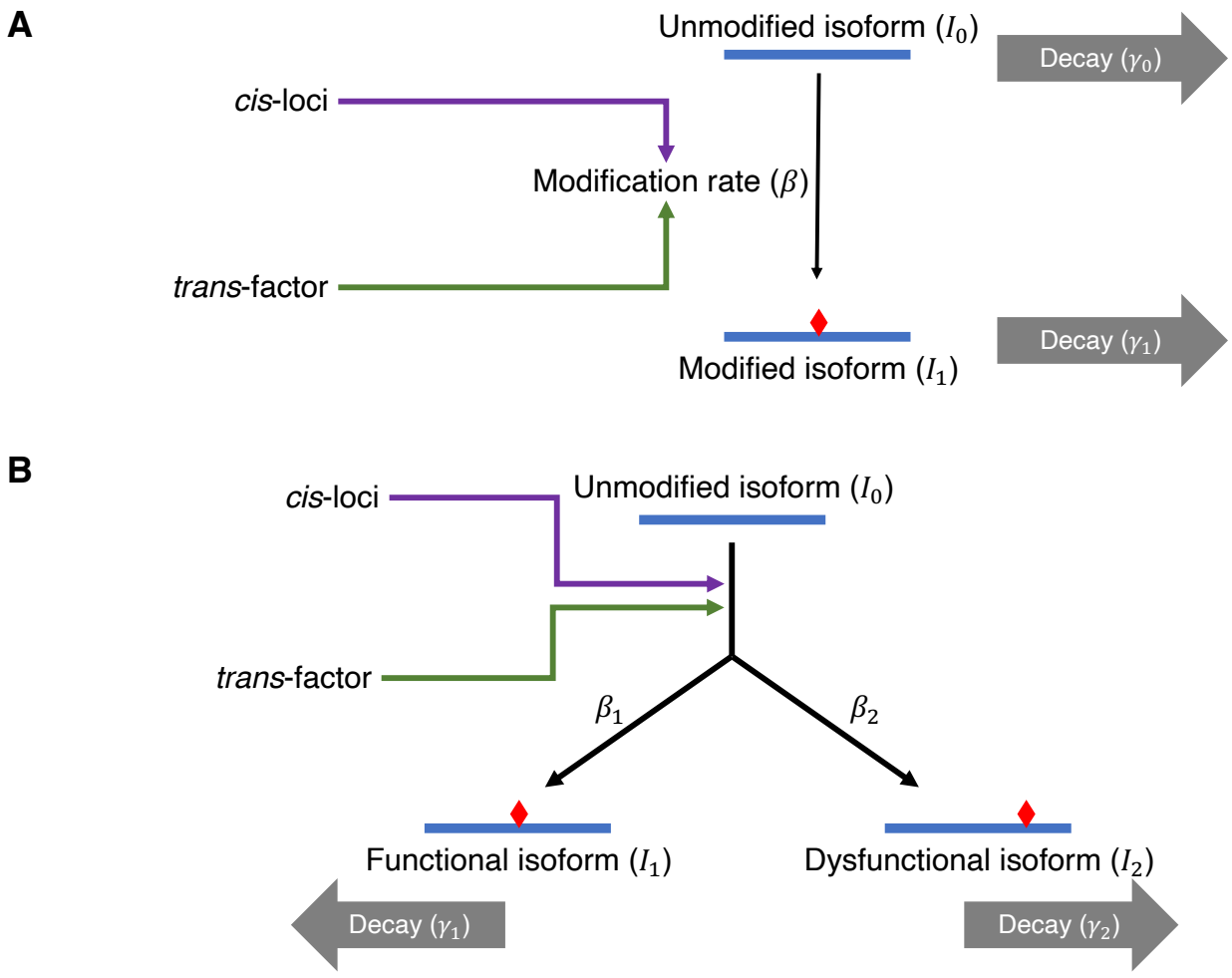

Figure 1: A schematic illustration of editing-type (A) and splicing-type (B) gene product diversity. (A) An unmodified isoform (I_0) is enzymatically converted to a modified isoform (I_1). The net per-molecule conversion rate (β) is determined jointly by a *trans*-factor (enzyme performing the modification process) and a set of *cis*-loci (sequence motif underlying affinity between enzyme and substrate). (B) The unmodified isoform I_0 can be converted into either a functional isoform (I_1) or a dysfunctional isoform (I_2) through the same modification process such that two conversion rates β_1 and β_2 are affected by the same *cis*-loci and *trans*-factor.

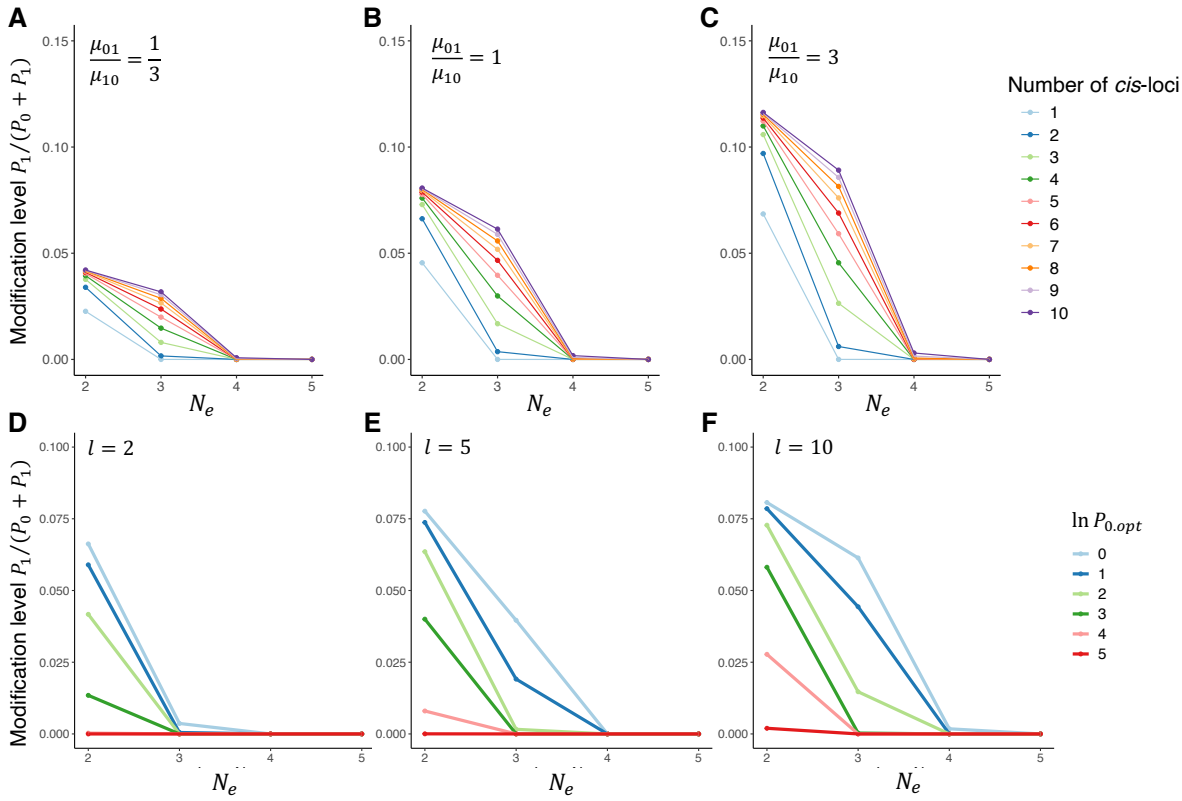

Figure 2: Scaling between mean modification level of a deleterious editing-type modification to effective population size N_e (shown in log10 scale). (A-C) Response of mean modification level to N_e given different combinations of *cis*-loci number (l) and mutation rates (μ_{01}, μ_{10}), with optimal expression level $P_{0.opt} = \exp(1)$ ($\ln P_{0.opt} = 1$). (A) Mutational bias is towards the null allele that does not facilitate modification. (B) Mutations of two directions have equal mutation rates. (C) Mutational bias is towards the effector allele that facilitates modification. (D-F) Response of mean modification level to N_e given different $P_{0.opt}$ with $l = 2$ (D), $l = 5$ (E), and $l = 10$ (F) in the absence of mutational bias. All results are derived with initial *cis*-genotypic value $v_0 = 0$, time of evolution $T = 10^8$ time steps, total mutation rate per *cis*-locus $\mu = \mu_{01} + \mu_{10} = 2 \times 10^{-9}$, $Q = 1$, $\gamma_0 = 1$, and $\gamma_1 = 1$. Optimal expression level $P_{0.opt}$ is set to be equal to P_0 in the absence of modification (i.e., $P_{0.opt} = \frac{\alpha}{\gamma_0}$) in all cases.

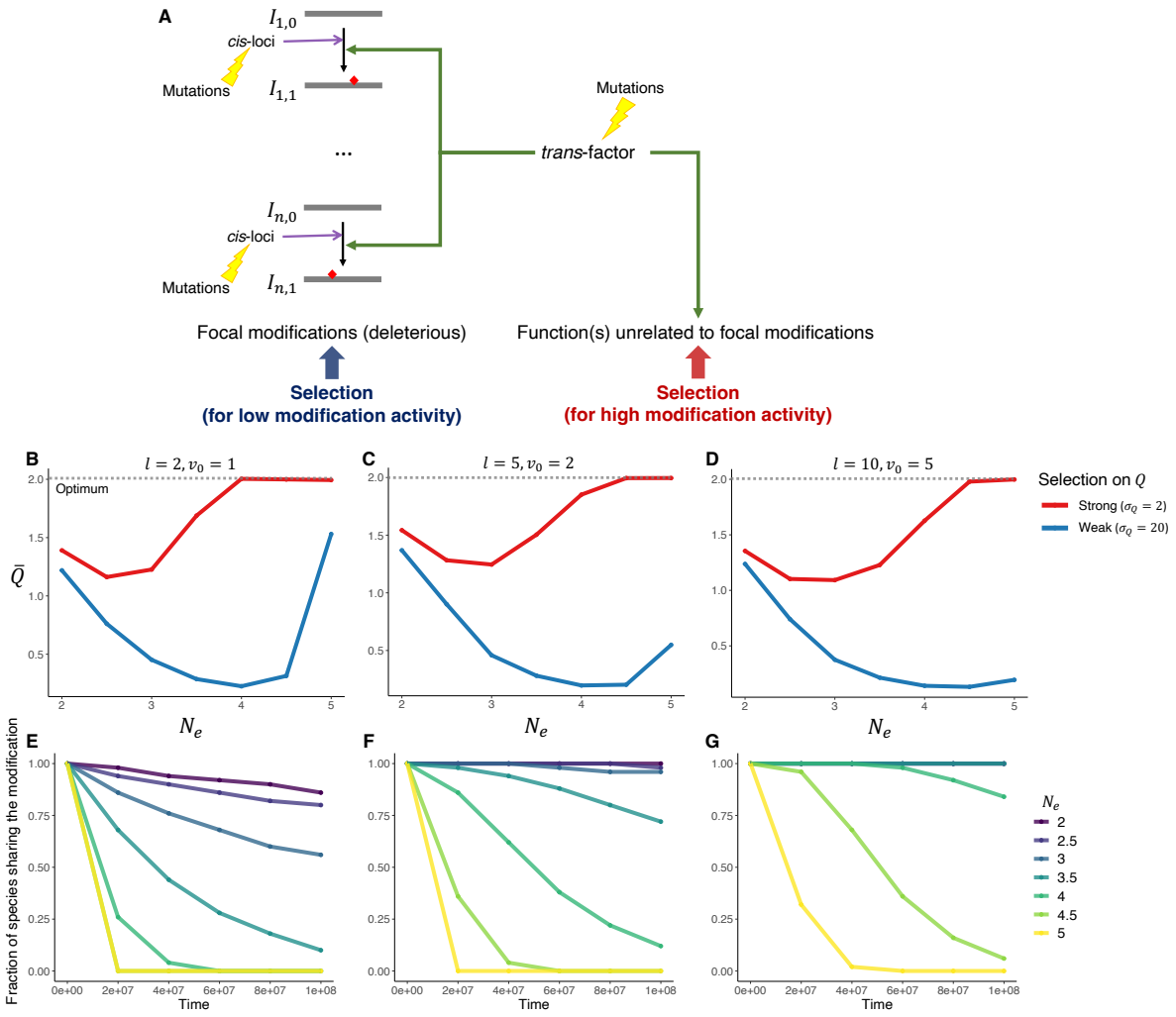

Figure 3: Coevolution of *cis*- and *trans*-acting loci when the gene product modification machinery is under opposing selection forces. (A) Schematic illustration of the scenario. The *trans*-factor, while causing a number of deleterious editing-type modification events (focal modifications), also performs an essential function independent of the focal modifications. Selection against deleterious modification may act to reduce the *trans*-genotypic value (Q), while selection mediated by the other function(s) act to maintain an optimal value of Q (Q_{opt}). (B-D) Non-monotonic response of mean of Q across lineages to N_e (shown in log10 scale) with Q under stabilizing selection and 100 genes subject to deleterious modification. Curves of different colors correspond to scenarios of strong (red) and weak (blue) selection on Q. Optimum of Q is denoted by the dashed line. All simulations started with an intermediate *cis*-genotypic value with largest corresponding genotypic space. (E-G) Sharing of modification events over time. Y-axes represent among-gene median of proportion of lineages (species) that share a modification event when selection on Q is strong ($\sigma_Q = 2$). When two curves in the same panel completely overlap, the one with the largest corresponding N_e is shown. In (B) and (E), $l = 2$ and $v_0 = 1$; in (C) and (F), $l = 5$ and $v_0 = 2$; in (D) and (G), $l = 10$ and $v_0 = 5$.

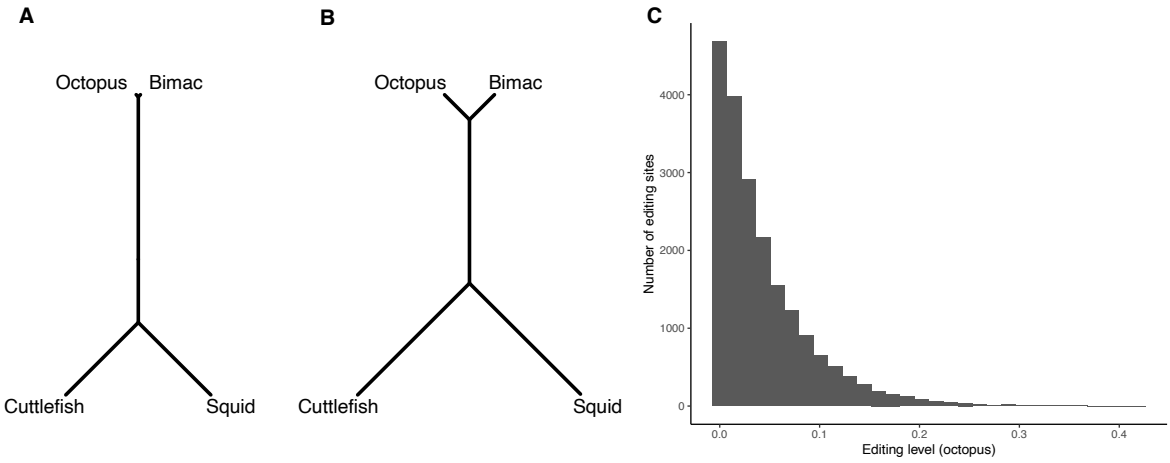

Figure 4: Evolutionary simulations recapitulated patterns of A-to-I RNA editing in four coleoid species, the octopus (*Octopus vulgaris*), the bimac (*O. bimaculoides*), the squid (*Doryteuthis pealeii*), and the cuttlefish (*Sepia oficianalis*). (A) Phylogenetic tree of four coleoid species. (B) Neighbor-joining tree of four coleoid species based on simulated editing levels. An unrooted version is shown in (A) as it is readily comparable to (B). (C) Distribution of editing levels across genes in the octopus.

Supplementary materials

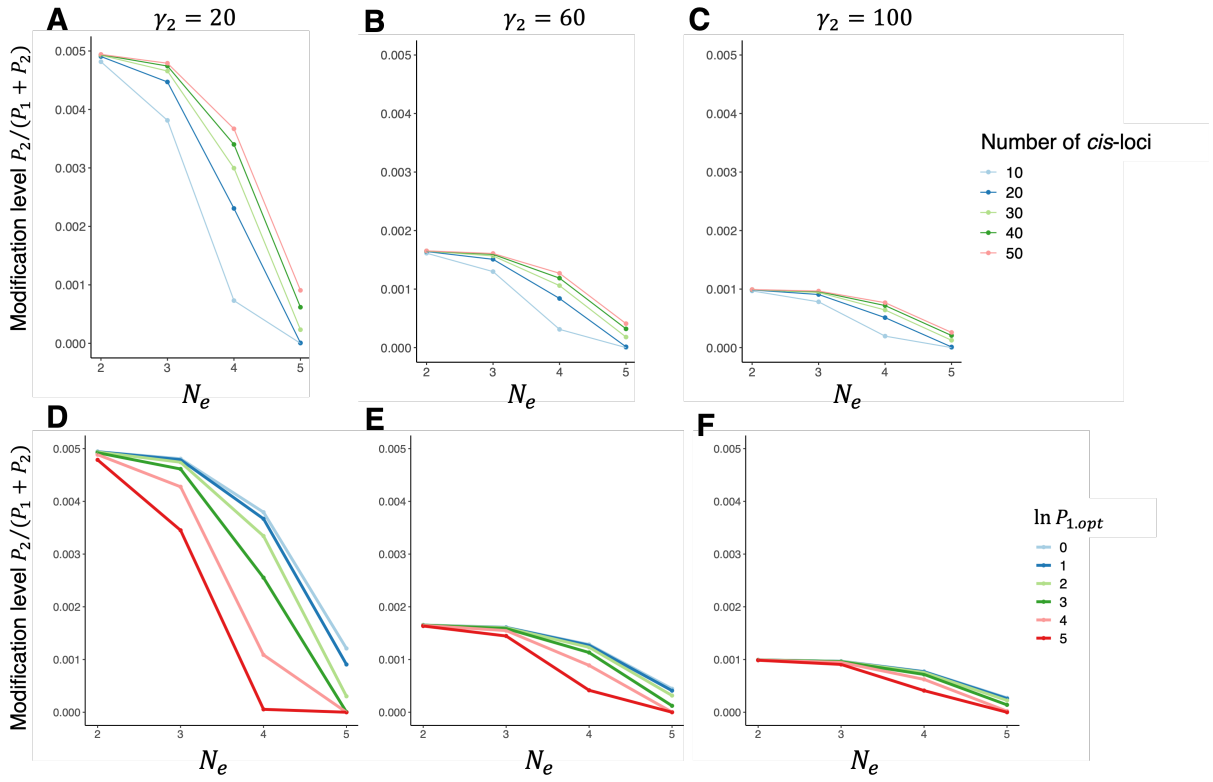

Figure S1: Scaling between mean modification level of splicing-type modification and effective population size N_e (shown in log10 scale). (A-C) Response of mean modification level to N_e under different combinations of *cis*-loci number (l) and decay rates of the dysfunctional isoform (γ_2), with with optimal expression level $P_{1.opt} = \exp(1)$ ($\ln P_{1.opt} = 1$). (D-F) Response of mean modification level to N_e under different $P_{1.opt} = \frac{\alpha}{\gamma_1}$ and γ_2, with $l = 50$. All results are derived with initial *cis*-genotypic value $v_0 = l$ (i.e., maximizing β_1 and minimizing β_2), time of evolution $T = 10^8$ time steps, $\mu_{01} = \mu_{10} = 10^{-8}$, $Q = 100$, $\gamma_0 = 0$, $\gamma_1 = 1$, and $P_{1.opt} = \frac{\alpha}{\gamma_1}$.

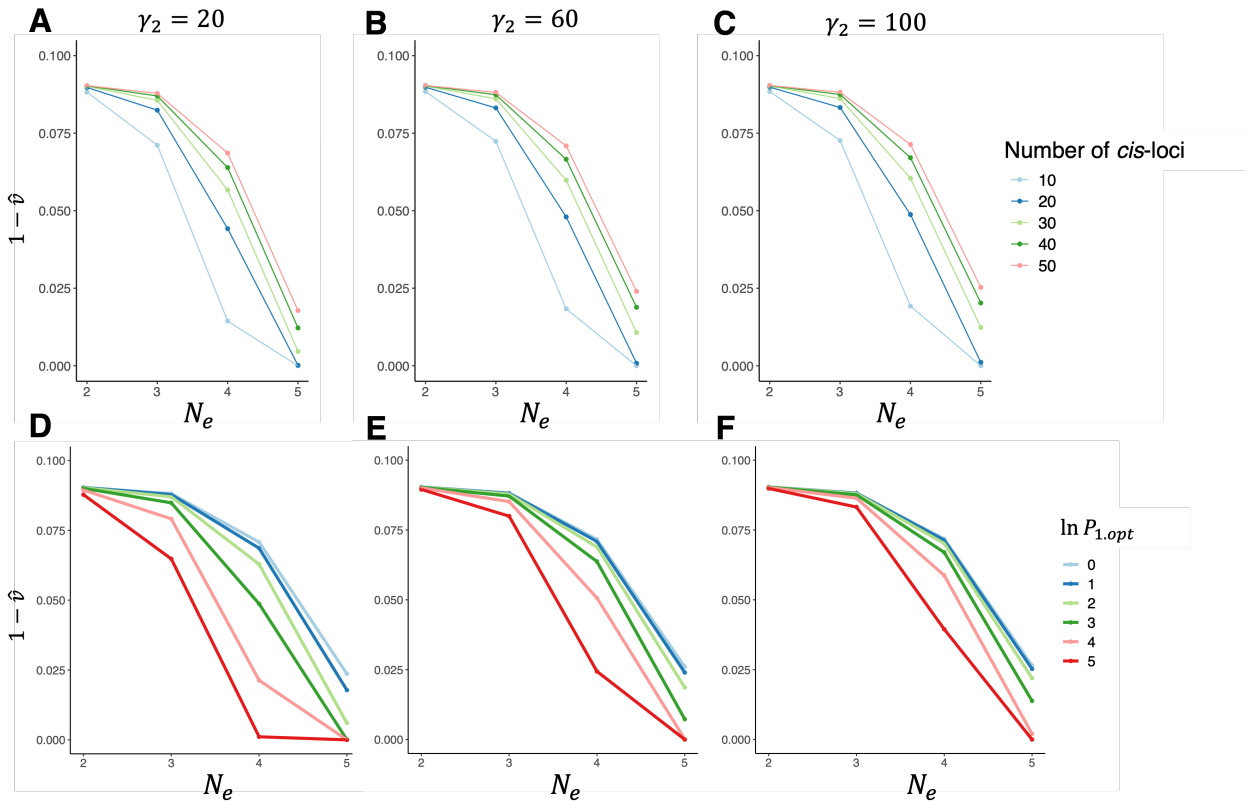

Figure S2: Scaling between normalized mean *cis*-genotypic value of splicing-type modification and N_e (shown in log10 scale). Represented by the Y-axes is $1-\hat{v}$, which reflects the degree to which *cis*-genotype favors production of the dysfunction and toxic isoform I_2. (A-C) Response of $1-\hat{v}$ to N_e under different combinations of l and γ_2, with optimal expression level $P_{1.opt} = \exp(1)$ ($\ln P_{1.opt} = 1$). (D-F) Response of $1-\hat{v}$ to N_e under different $P_{1.opt}$ and γ_2, with $l = 50$. All results are derived with initial *cis*-genotypic value $v_0 = l$, time of evolution $T = 10^8$ time steps, and $\mu_{01} = \mu_{10} = 10^{-8}$, $Q = 100$, $\gamma_0 = 0$, $\gamma_1 = 1$, and $P_{1.opt} = \frac{\alpha}{\gamma_1}$.

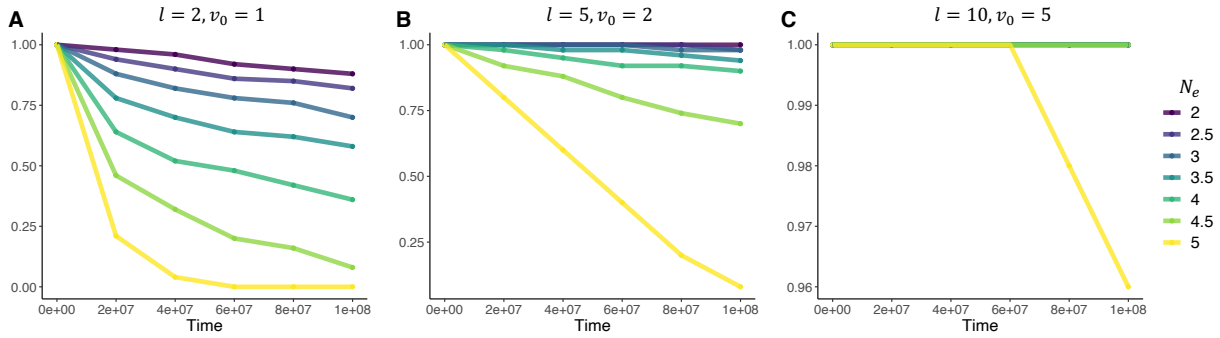

Figure S3: Sharing of modification events over time. Y-axes represent among-gene median of proportion of lineages (species) that share a modification event when selection on Q is weak ($\sigma_Q = 20$). When two curves in the same panel completely overlap, the one with the largest corresponding N_e is shown.

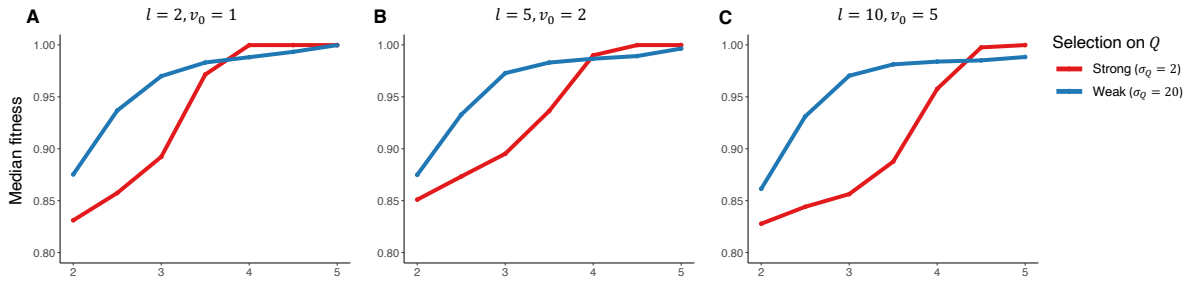

Figure S4: Response of median fitness across lineages (species) to N_e when the modification process is under opposing selection forces. (A) $l = 2$, $v_0 = 1$. (B) $l = 5$, $v_0 = 2$. (C) $l = 10$, $v_0 = 5$. Maximum fitness (fitness value at the global optimum) is equal to 1.

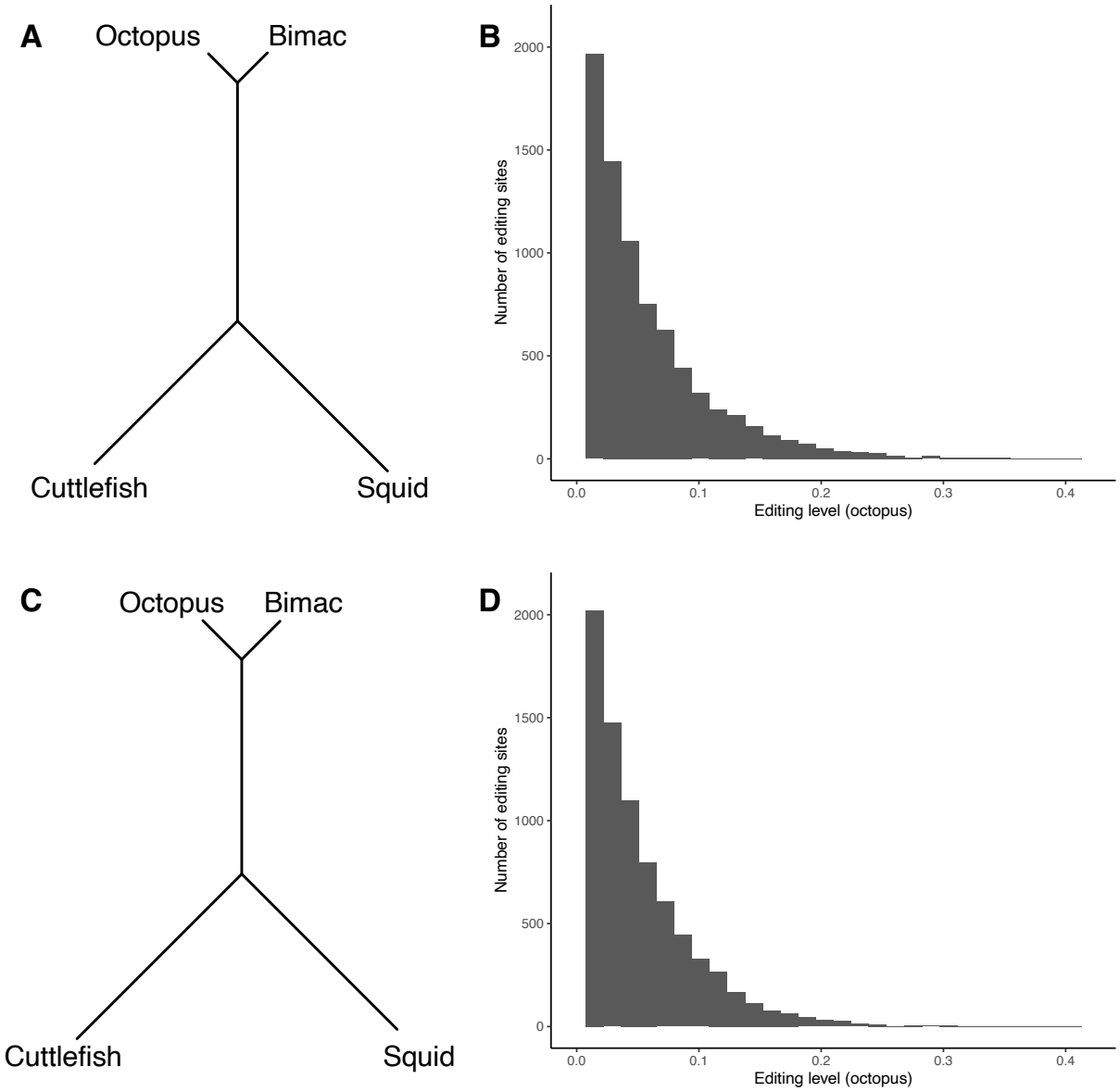

Figure S5: (A) Neighbor-joining tree of four coleoid species based on simulated neutral editing levels. (B) Distribution of neutral editing levels in the octopus. (C) Neighbor-joining tree of four coleoid species based on simulated deleterious editing levels. (D) Distribution of deleterious editing levels in the octopus.